



RT09 Conference Speech Recognition System

Dr. Veton Këpuska
Arthur Kunkle



LVSR System Goals

- **Facilitate re-use by identifying existing solutions for speech recognition tasks**
 - HTK (HMM Toolkit)
 - SRILM (SRI Language Modeling Toolkit)
- **Use a robust “profile” mechanism to track unique sets of experiment input parameters and organize all models and evaluation results in a single area.**
- **Ingest data from a variety of sources and account for format and transcription disparities.**
- **Provide System performance metrics throughout different processing stages for system improvement and research purposes.**

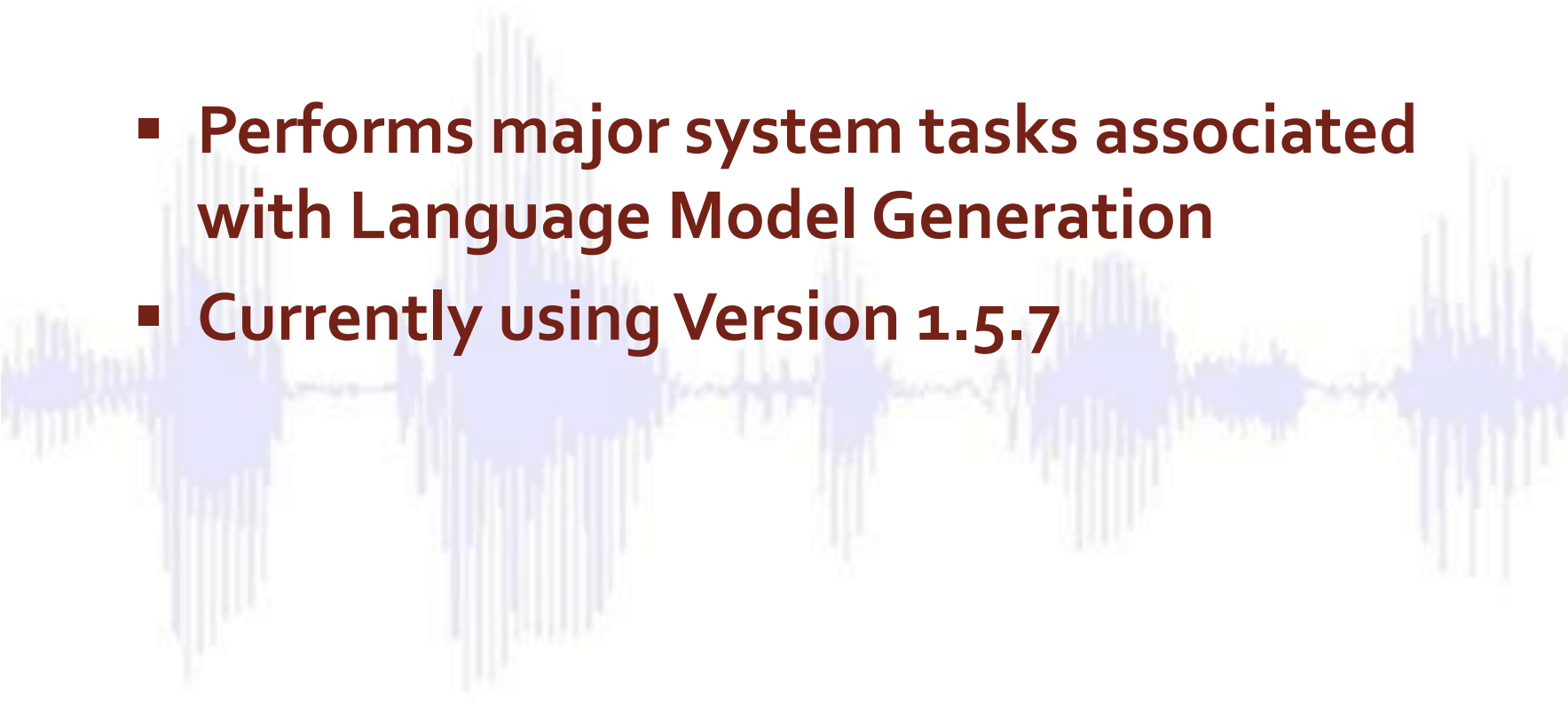
HTK System Interface

- Performs major system tasks associated with corpora data preparation, acoustic modeling, Viterbi Search, and lattice rescoring
- Selected for well-organized baseline. Easy to extend for future goals.
- Currently using Version 3.4.1

HTK Tool	Purpose
HCopy	copy one or more data files to a designated output file, optionally converting the data into a parameterized form
HDMan	used to prepare a pronunciation dictionary from one or more sources.
HLRescore	HLRescore is a general lattice post-processing tool. It reads lattices (for example produced by HVite) and finds the best path.
HVite	HVite is a general-purpose Viterbi word recognizer. It will match a speech file against a network of HMMs and output a transcription for each. When performing N-best recognition a word level lattice containing multiple hypotheses can also be produced
HERest	perform a single re-estimation of the parameters of a set of HMMs, or linear transforms, using an <i>embedded training version of the Baum-Welch algorithm</i>

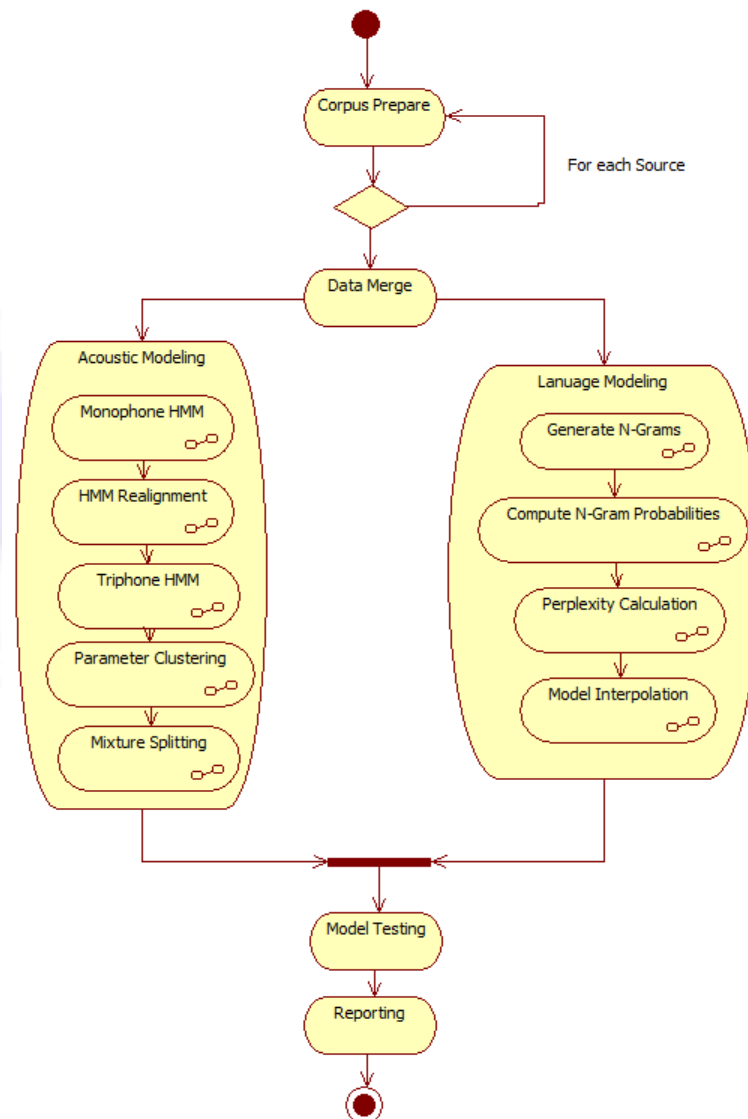


SRILM System Interface

- 
- **Performs major system tasks associated with Language Model Generation**
 - **Currently using Version 1.5.7**

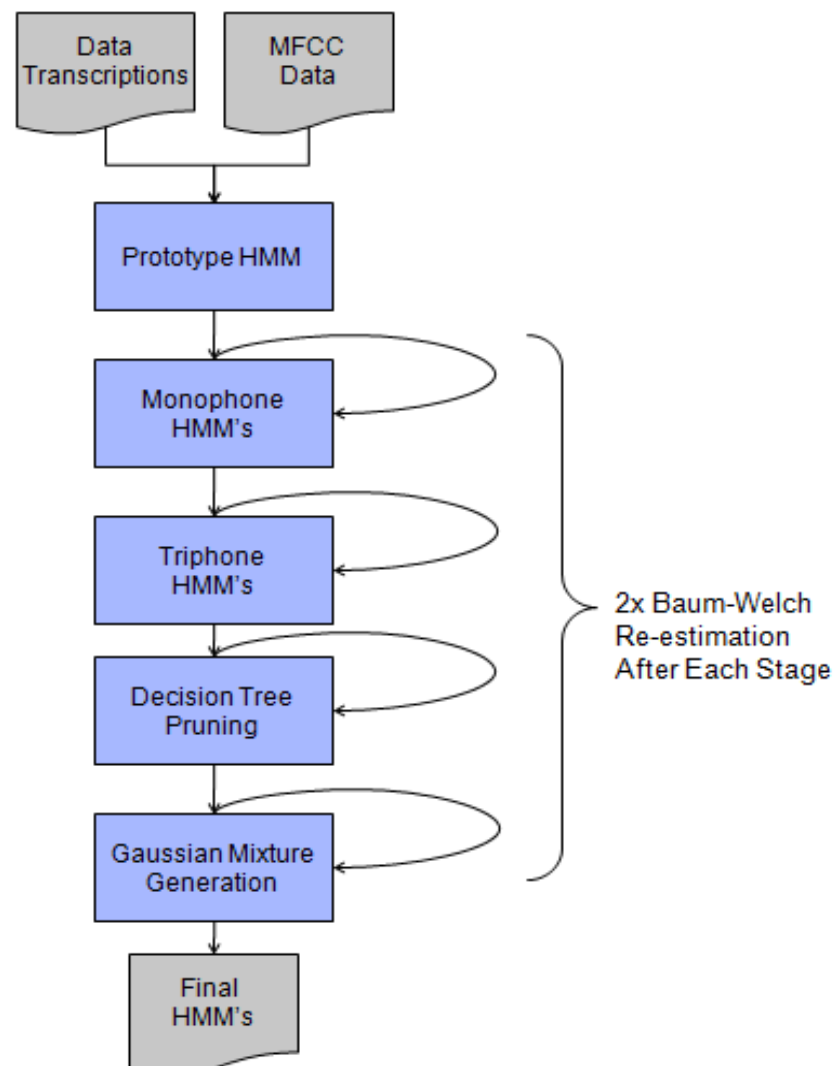
LVSR System Overview

- **Designed to be an end-to-end solution for:**
 - Processing labeled speech data from various sources
 - train acoustic and language models
 - perform speech-to-text recognition
- **Perl used extensively to abstract HTK calls and provide transcription processing.**



Acoustic Model Training

- Perl code manages the workflow between HTK Acoustic Modeling tools.
- Builds phone-based HMM models iteratively:
 - 3-state monophones
 - Context-dependent Triphones
 - Tied-State Triphones using Decision Trees
 - Multi-Gaussian mixture Triphones



Language Model Training Dev. Testing

Train Data : RT0607_TRAIN	Test Data: RT0607_TRAIN		
N-gram Order	5		
Training Parameters	ppl	ppl1	logprob
kndiscount_kn-modify-counts-at-end	70.17	74.49	-218926
kndiscount	55.79	59.03	-207111
ukndiscount	49.35	30.05	-200800
ukndiscount_kn-modify-counts-at-end	28.67	30.05	-172828
default	7.69	7.91	-105041
interpolate	7.69	7.91	-105041
wbdiscout	7.56	7.78	-104172
ndiscount	6.79	6.97	-98620

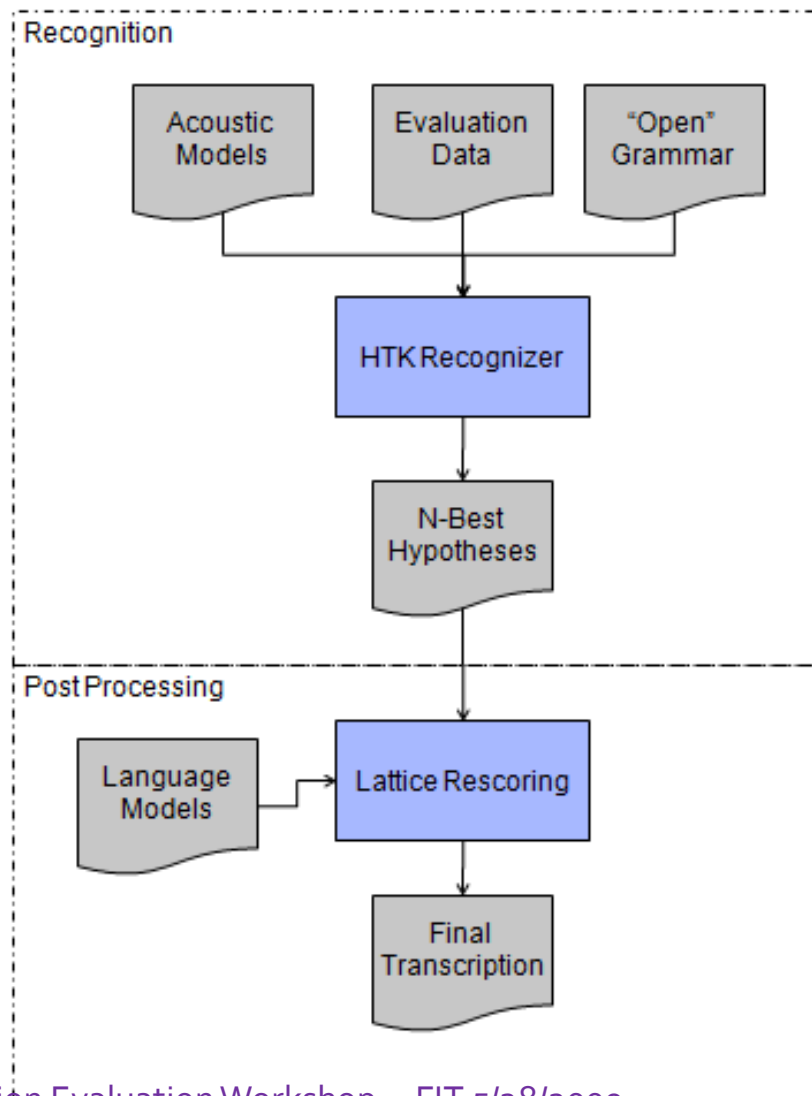
Train Data : TIMIT_TRAIN	Test Data: TIMIT_TRAIN		
N-gram Order	5		
Training Parameters	ppl	ppl1	logprob
kndiscount_kn-modify-counts-at-end			
kndiscount			
ukndiscount	53.75	83.00	-83711
ukndiscount_kn-modify-counts-at-end	49.02	74.94	-81777
default	5.99	7.28	-37604
interpolate	5.99	7.28	-37604
wbdiscout	4.70	5.57	-32536
ndiscount	4.50	5.30	-31579

Train Data : RT0607_TRAIN	Test Data: RT0607_TEST		
N-gram Order	5		
Training Parameters	ppl	ppl1	logprob
kndiscount_kn-modify-counts-at-end	71.37	74.51	-18932
kndiscount	55.47	57.76	-17814
ukndiscount	44.84	46.59	-16870
ukndiscount_kn-modify-counts-at-end	25.14	25.97	-14303
default	48.83	50.78	-17248
interpolate	48.83	50.78	-17248
wbdiscout	44.43	46.16	-16829
ndiscount	42.34	43.97	-16615

Train Data : TIMIT_TRAIN	Test Data: TIMIT_TEST		
N-gram Order	5		
Training Parameters	ppl	ppl1	logprob
kndiscount_kn-modify-counts-at-end			
kndiscount			
ukndiscount	9.17	11.84	-4735
ukndiscount_kn-modify-counts-at-end	9.13	11.80	-4727
default	11.74	15.61	-5264
interpolate	11.74	15.61	-5264
wbdiscout	13.65	10.41	-5007
ndiscount	10.47	13.73	-5019

Recognizer and Post-Processing

- Utilizes the HTK Recognizer HVite to create N-best hypothesis lattices for evaluation data against an “open” grammar network (no Language Model)
- Language Models from SRILM then applied to N-best lattices to generate final hypothesis.
- Other factors such as Insertion Penalties, Grammar/Acoustic Scaling Factors applied.
- Results are then displayed using HTK tools or NIST scoring software (SCTK)



RTog Evaluation Experiment Setup

- **Model training was performed in advance of the evaluation.**
- **Acoustic Models are built iteratively**
- **Dictionary based upon CMU Pronunciation Dictionary**

Experiment Parameter	Value
Final Triphone HMM Count	8422
Dictionary Count	16426
Language Model Order	3
Language Model Discounting	Kneser-Ney
Acoustic Modeling Corpora Used	TIMIT (initial) AMI RT06 RT07
Language Modeling Corpora Used	RT06 RT07
MFCC Coefficients	39 (13+13+13)
Gaussian Mixtures Used	9 (incremented by 3)

RTog Evaluation Results

- FIT participated in the *speech-to-text* RTog task, which involves generating time-stamped transcriptions of spoken word data.
- Evaluation data is 180 minutes of data collected from ten meetings occurring at two facilities.
- FIT processed Individual Head Microphone data with reference segmentations

Regions	Total Time	Total Words	Correct	Substitutions	Insertions	Deletions	Word Error	Sentence Error
11852	55883.9	42247	11016 (26.1%)	29455 (69.7%)	15981 (37.8%)	1176 (4.2%)	47212 (111.8%)	(50.8%)



LVSR Improvement Goals

- **Clearly room for improvement in system performance**
- **Following areas are planned for additional analysis in the near term:**
 - Determine Cause of Excessive Insertion Rate
 - Poor Data Availability for a Large Percentage of Triphone Models
 - Improving User Interface
 - Knowledge Depth of the HTK and SRILM Utilities

Research Projects

■ Rich Transcription Evaluation

- Baseline Performance – note yet complete
- Augment HTK with Wake-Up-Word (WUW) Technology (next task)
 - Performance of WUW speech recognition system on whole word recognition task compared against leading commercial speech recognition system (Microsoft's SDK 5.1) showed **from 919.67% to 1760.00%** and against the leading academic speech recognition system HTK **from 1449.75% to 15166.15% improvement (measured as relative error rate reduction)**

Research Projects

- **Wake-Up-Word (WUW) Speech Recognition (SR)**
 - “You talkin’ to me?” - Key to *spontaneous and natural* human-machine-interaction is to reliably determine if a user is addressing the computer .
 - An WUW is a SR system that is capable of reliably determining if the user is addressing the computer.
- **Alerting vs. Referential Context**
 - Alerting:
 - “**Computer!** Begin PowerPoint Presentation”
 - Referential:
 - “I have a **computer** with dual core 64 bit Intel Processor and 4 GB of RAM.”
 - On going creative data collection and analysis procedure based on DVD movies to facilitate and refine our understanding of alerting vs . referential context using only prosodic features.
- **Development**
 - PowerPoint Commander – Voice Activated and Controlled PowerPoint Presentation
 - Elevator Simulator



**Robert DeNiro in
“Taxi Driver”**



Questions?

